Multiagent Value Iteration in Markov Games

Amy Greenwald
Brown University

with
Michael Littman and Martin Zinkevich

Stony Brook Game Theory Festival
July 21, 2005

Agenda

Theorem

Value iteration converges to a stationary optimal policy in Markov decision processes.

Question

Does multiagent value iteration converge to a stationary equilibrium policy in Markov games?

Multiagent Q-Learning

Minimax-Q Learning [Littman 1994]

- provably converges to stationary minimax equilibrium policies in zero-sum Markov games

Nash-Q Learning [Hu and Wellman 1998]
Correlated-Q Learning [G and Hall 2003]

- converge empiricially to stationary equilibrium policies on a testbed of general-sum Markov games

Multiagent Value Iteration $\longrightarrow$ Cyclic Equilibria

Theory

Multiagent value iteration converges to cyclic equilibrium policies in Marty's game.

Experiments

Multiagent value iteration converges to cyclic equilibrium policies

- Michael's game
- randomly generated Markov games
- Grid Game 1 [Hu and Wellman 1998]
- Shopbots and Pricebots [G and Kephart 1999]

Markov Decision Processes (MDPs)

Decision Process

- S is a set of states
- A is a set of actions
- $R : S \times A \to \mathbb{R}$ is a reward function
- $P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

MDP = Decision Process + Markov Property:

$$P[s_{t+1} \mid s_t, a_t, \ldots, s_0, a_0] = P[s_{t+1} \mid s_t, a_t]$$

$\forall t,\ \forall s_0, \ldots, s_t \in S,\ \forall a_0, \ldots, a_t \in A$

Bellman's Equations

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P[s' \mid s,a] V^*(s') \quad (1)$$

$$V^*(s) = \max_{a \in A} Q^*(s,a) \quad (2)$$

Value Iteration

VI(MDP, γ)

Inputs	discount factor γ
Output	optimal state-value function V^*
	optimal action-value function Q^*
Initialize	V arbitrarily

```
REPEAT
    for all s ∈ S
        for all a ∈ A
            Q(s,a) = R(s,a) + γ Σ_{s'} P[s' | s,a] V(s')
        V(s) = max_a Q(s,a)
FOREVER
```

Markov Games

Stochastic Game

- N is a set of players
- S is a set of states
- A_i is the ith player's set of actions
- $R_i(s, \vec{a})$ is the ith player's reward at state s given action vector $\vec{a}$
- $P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0]$ is a probabilistic transition function that describes transitions between states, conditioned on past states and actions

Markov Game = Stochastic Game + Markov Property:

$$P[s_{t+1} \mid s_t, \vec{a}_t, \ldots, s_0, \vec{a}_0] = P[s_{t+1} \mid s_t, \vec{a}_t]$$

$\forall t$, $\forall s_0, \ldots, s_t \in S$, $\forall \vec{a}_0, \ldots, \vec{a}_t \in A$

Bellman's Analogue

$$Q_i^*(s,\vec{a}) = R_i(s,\vec{a}) + \gamma \sum_{s'} P[s' \mid s,\vec{a}] V_i^*(s') \tag{3}$$

$$V_i^*(s) = \sum_{\vec{a}\in A} \pi^*(s,\vec{a}) Q_i^*(s,\vec{a}) \tag{4}$$

Foe-VI	$\pi^*(s) = (\sigma_1^*, \sigma_2^*)$, a minimax equilibrium policy [Shapley 1953, Littman 1994]
Friend-VI	$\pi^*(s) = e_{\vec{a}^*}$ where $\vec{a}^* \in \arg\max_{\vec{a}\in A} Q_i^*(s,\vec{a})$ [Littman 2001]
Nash-VI	$\pi^*(s) \in \text{Nash}(Q_1^*(s), \ldots, Q_n^*(s))$ [Hu and Wellman 1998]
CE-VI	$\pi^*(s) \in \text{CE}(Q_1^*(s), \ldots, Q_n^*(s))$ [G and Hall 2003]

Multiagent Value Iteration

MULTI–VI(MGame, γ, f)

Inputs	discount factor γ
	selection mechanism f
Output	equilibrium state-value function V^*
	equilibrium action-value function Q^*
Initialize	V arbitrarily

```
REPEAT
    for all s ∈ S
        for all a⃗ ∈ A
            for all i ∈ N
                Q_i(s, a⃗) = R_i(s, a⃗) + γ Σ_{s'} P[s' | s, a⃗] V_i(s')
        π(s) ∈ f(Q_1(s), ..., Q_n(s))
        for all i ∈ N
            V_i(s) = Σ_{a⃗∈A} π(s, a⃗) Q_i(s, a⃗)
FOREVER
```

Friend-or-Foe-VI *always* converges [Littman 2001]

Nash-VI and CE-VI converge to stationary equilibrium policies in zero-sum & common-interest Markov games [GZ and Hall 2005]

Cyclic Correlated Equilibria

A cyclic policy ρ is a sequence of $k < \infty$ stationary policies.

$$V_i^{\rho,t}(s) = \sum_{\vec{a} \in A} \rho_t(s, \vec{a}) Q_i^{\rho,t}(s, \vec{a}) \tag{5}$$

$$Q_i^{\rho,t}(s, \vec{a}) = R_i(s, \vec{a}) + \gamma \sum_{s' \in S} P[s' \mid s, \vec{a}] V_i^{\rho, t \bmod k+1}(s') \tag{6}$$

A cyclic policy of length k is a correlated equilibrium if for all $i \in N$, $s \in S$, $a_i' \in A_i$, and $t \in \{1, \ldots, k\}$,

$$\sum_{\vec{a}_{-i} \in A_{-i}} \rho_t(s, \vec{a}_{-i} \mid a_i) Q_i^{\rho,t}(s, \vec{a}_{-i}, a_i) \geq \sum_{\vec{a}_{-i} \in A_{-i}} \rho_t(s, \vec{a}_{-i} \mid a_i) Q_i^{\rho,t}(s, \vec{a}_{-i}, a_i') \tag{7}$$

Michael's Game: Best-Response Cycle

Observation

Michael's game has no stationary deterministic equilibrium policy when $\gamma > \frac{1}{2}$.

Proof

(A quits, B quits)	$\Rightarrow$	A prefers send to quit ($2\gamma > 1$)
(A sends, B quits)	$\Rightarrow$	B prefers send to quit ($0 > -1$)
(A sends, B sends)	$\Rightarrow$	A prefers quit to send ($1 > 0$)
(A quits, B sends)	$\Rightarrow$	B prefers quit to send ($-1 > -2$)

Michael's Game: Cyclic Policy

Observation

Michael's game has a deterministic cyclic equilibrium policy when $\gamma = \frac{2}{3}$.

Example

	Policy	$V(A)$	$V(B)$
1	(A quits, B sends)	$(1, -2)$	$\left(\frac{8}{9}, -\frac{4}{9}\right)$
2	(A sends, B sends)	$\left(\frac{4}{3}, -\frac{2}{3}\right)$	$\left(\frac{8}{9}, -\frac{4}{9}\right)$
3	(A sends, B quits)	$\left(\frac{4}{3}, -\frac{2}{3}\right)$	$(2, -1)$
4	(A quits, B quits)	$(1, -2)$	$(2, -1)$

Michael's Game: Equilibrium Constraints

$$V_A^1(A) = Q_A^1(A, quit) = 1 > \frac{16}{27} = 0 + \left(\frac{2}{3}\right)\left(\frac{8}{9}\right) = 0 + \gamma V_A^2(B) = Q_A^1(A, send)$$

$$V_A^2(A) = Q_A^2(A, send) = 0 + \gamma V_A^3(B) = 0 + \left(\frac{2}{3}\right)(2) = \frac{4}{3} > 1 = Q_A^2(A, quit)$$

$$V_A^3(A) = Q_A^3(A, send) = 0 + \gamma V_A^4(B) = 0 + \left(\frac{2}{3}\right)(2) = \frac{4}{3} > 1 = Q_A^3(A, quit)$$

$$V_A^4(A) = Q_A^4(A, quit) = 1 > \frac{16}{27} = 0 + \left(\frac{2}{3}\right)\left(\frac{8}{9}\right) = 0 + \gamma V_A^1(B) = Q_A^4(A, send)$$

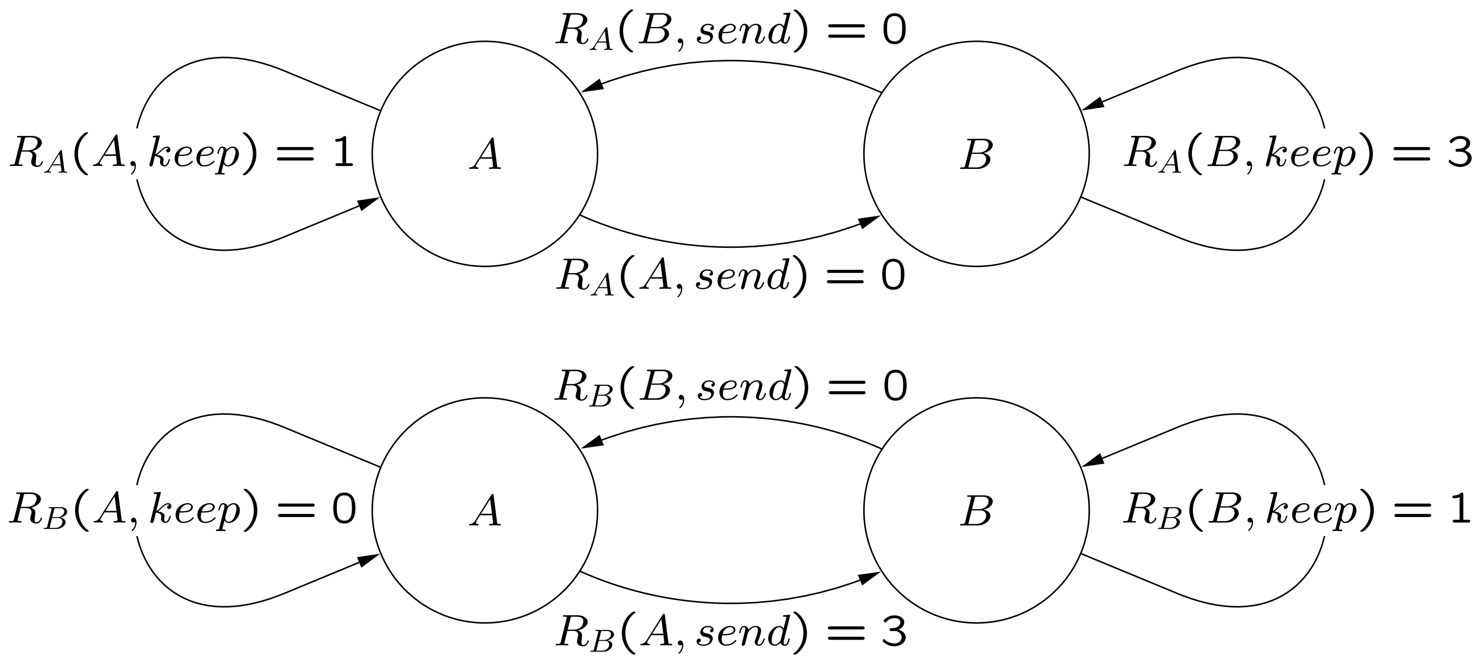
$$V_B^1(B) = Q_B^1(B, send) = 0 + \gamma V_B^2(A) = 0 + \left(\frac{2}{3}\right)\left(-\frac{2}{3}\right) = -\frac{4}{9} > -1 = Q_B^1(B, quit)$$

$$V_B^2(B) = Q_B^2(B, send) = 0 + \gamma V_B^3(A) = 0 + \left(\frac{2}{3}\right)\left(-\frac{2}{3}\right) = -\frac{4}{9} > -1 = Q_B^2(B, quit)$$

$$V_B^3(B) = Q_B^3(B, quit) = -1 > -\frac{4}{3} = 0 + \left(\frac{2}{3}\right)(2) = 0 + \gamma V_B^4(A) = Q_B^3(B, send)$$

$$V_B^4(B) = Q_B^4(B, quit) = -1 > -\frac{4}{3} = 0 + \left(\frac{2}{3}\right)(2) = 0 + \gamma V_B^1(A) = Q_B^4(B, send)$$

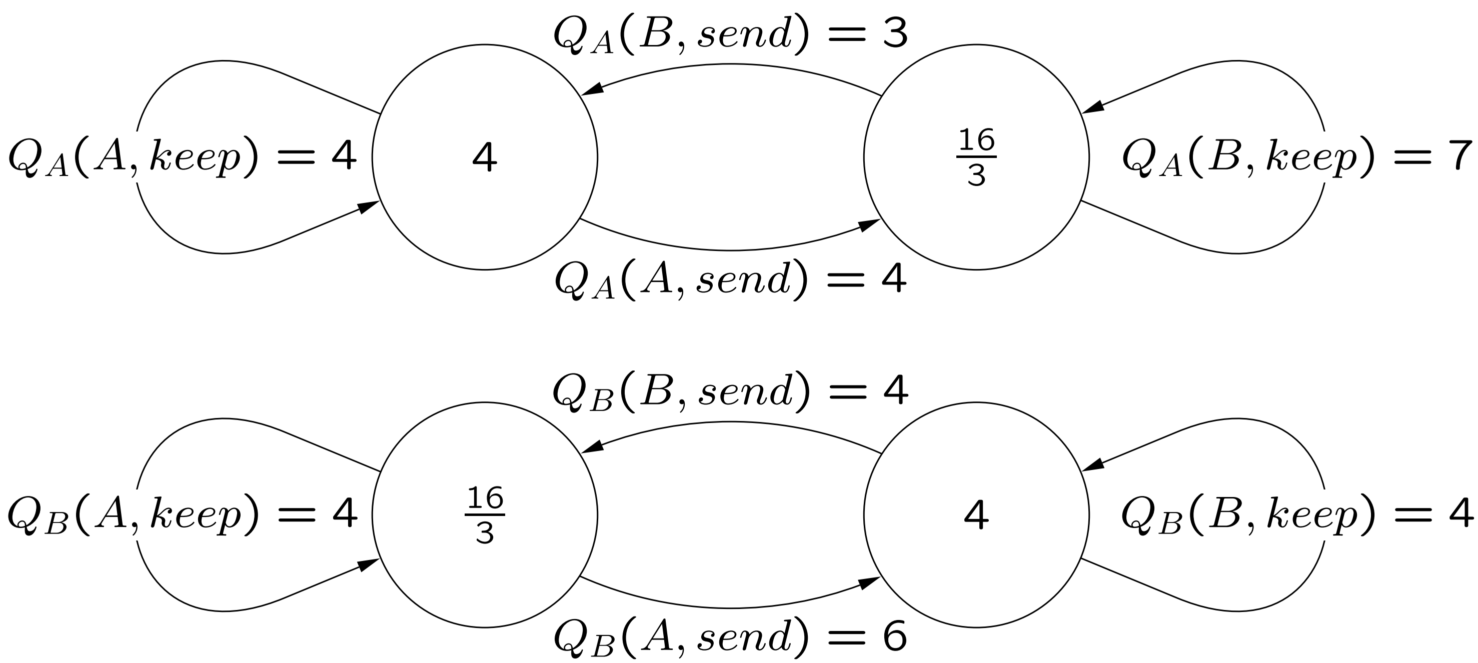
Marty's Game: Rewards



Observation [ZGL 2005]

Marty's game has no stationary deterministic equilibrium policy when $\gamma = \frac{3}{4}$.

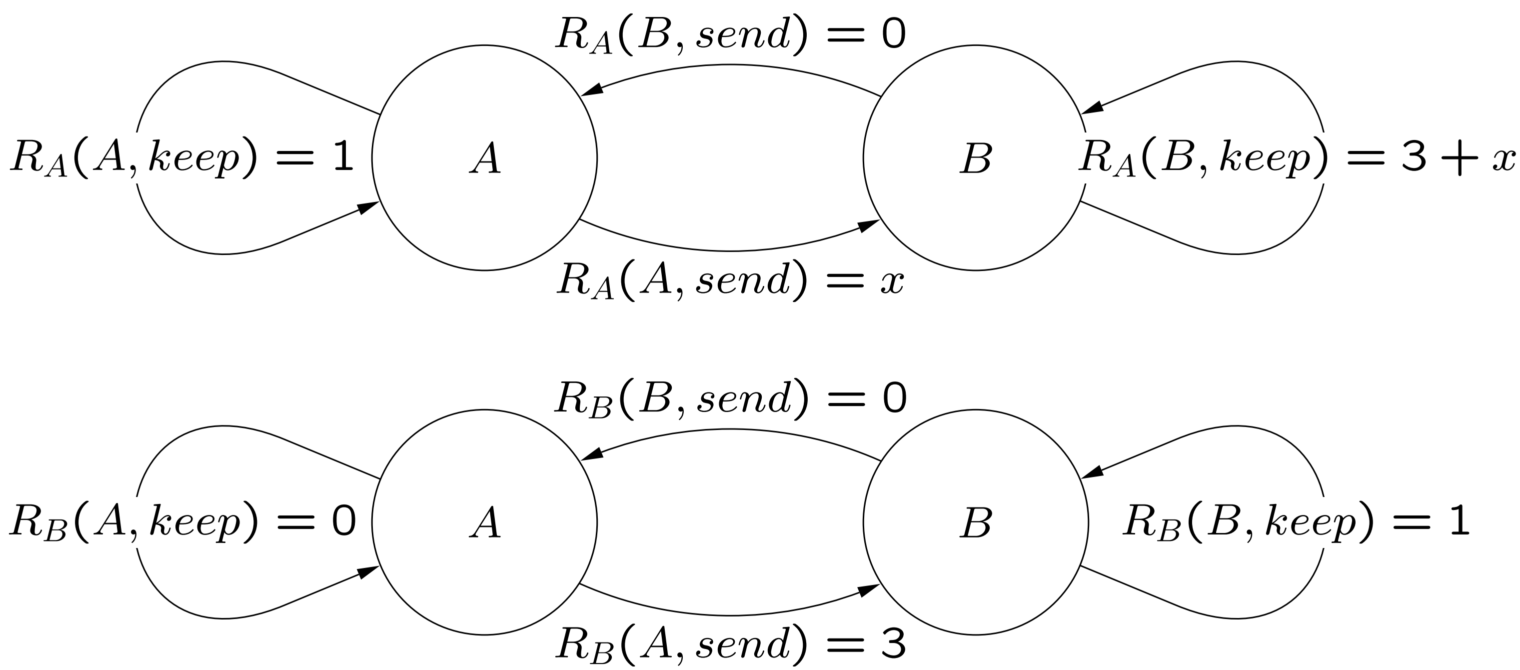
Marty's Game: Q-Values and Values



Theorem [ZGL 2005]

Marty's game has a unique (probabilistic) stationary equilibrium policy.

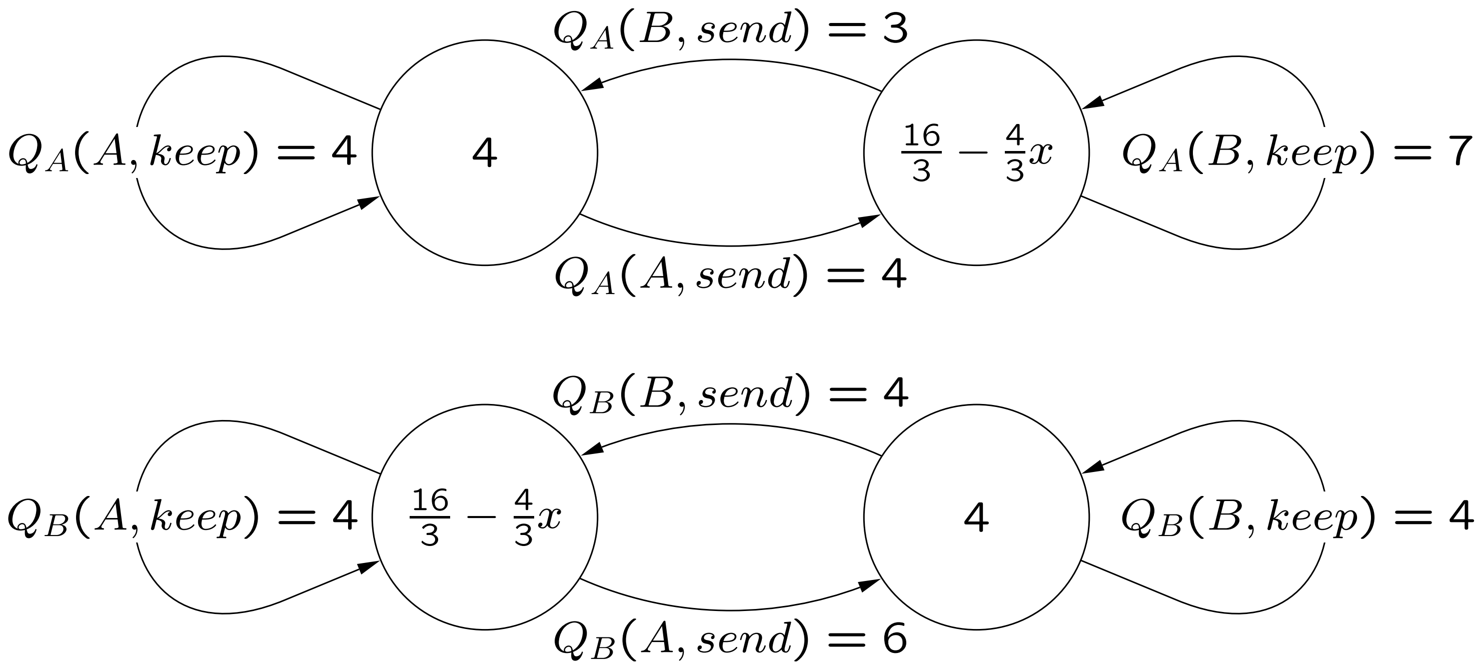
Marty's Games: Tweaked Rewards



Observation [ZGL 2005]

These games have no stationary deterministic equilibria for $-1 < x < \frac{7}{4}$ and $\gamma = \frac{3}{4}$.

Marty's Games: Q-Values and Tweaked Values



Theorem [ZGL 2005]

These games have unique (probabilistic) stationary equilibrium policies.

Negative Result

Theorem [ZGL 2005]

There exist an infinite number of Marty's games with the same Q-values, but different V-values and different stationary equilibrium policies.

Negative Result

Theorem [ZGL 2005]

There exist an infinite number of Marty's games with the same Q-values, but different V-values and different stationary equilibrium policies.

Positive Result

Theorem [ZGL 2005]

In Marty's games, given any "natural" equilibrium selection mechanism, there exists some $k > 1$ s.t. multiagent value iteration converges to a cyclic equilibrium policy of length k.

Random Markov Games

$|N| = 2$

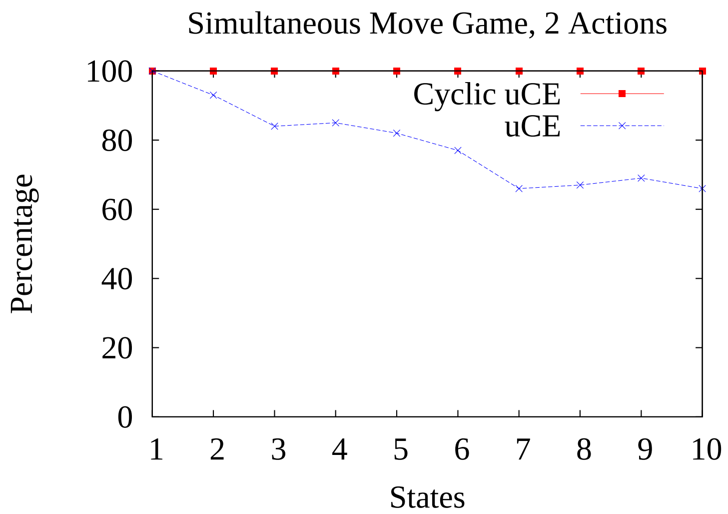
$|A| \in \{2, 3\}$

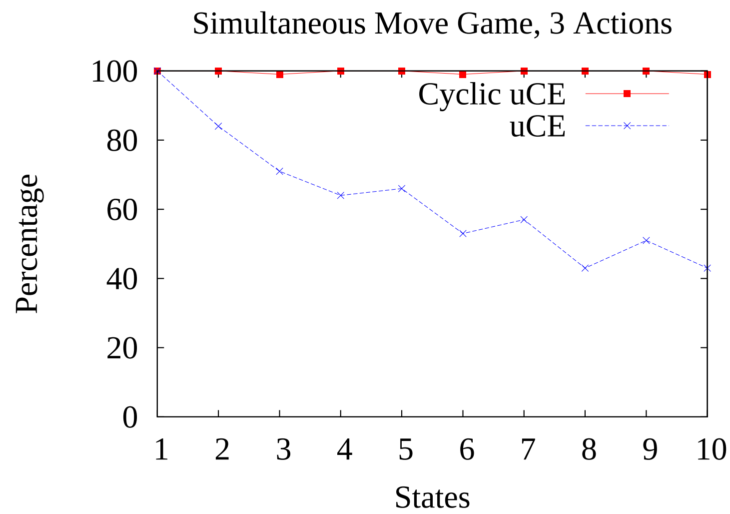
$|S| \in \{1, \ldots, 10\}$

Random Rewards $\in [0, 99]$

Random Deterministic Transitions

$\gamma = \frac{3}{4}$





Multiagent Value Iteration in Markov Games

Summary of Observations

- Multiagent value iteration converges to nonstationary deterministic cyclic equilibrium policies in Marty's and Michael's games.

- Multiagent value iteration converges empirically to not necessarily deterministic, not necessarily stationary, cyclic equilibrium policies in randomly generated deterministic Markov games.

Multiagent Value Iteration in Markov Games

Summary of Observations

- Multiagent value iteration converges to nonstationary deterministic cyclic equilibrium policies in Marty's and Michael's games.

- Multiagent value iteration converges empirically to not necessarily deterministic, not necessarily stationary, cyclic equilibrium policies in randomly generated deterministic Markov games.

Open Questions

- Do deterministic cyclic equilibrium policies necessarily exist in turn-taking games? If so, does multiagent value iteration necessarily converge to deterministic cyclic equilibrium policies in turn-taking games?

- Just as multiagent value iteration necessarily converges to stationary equilibrium policies in zero-sum Markov games, does multiagent value iteration necessarily converge to nonstationary cyclic equilibrium policies in general-sum Markov games?

The Answer is No!

Multiagent value iteration does not necessarily converge to stationary equilibrium policies in general-sum Markov games, regardless of the equilibrium selection mechanism.